

## What is $TAP-k$ ?

Here we refer to the measure defined by Carroll, H. D., Kann, M. G., Sheetlin, S. L., and Spouge, J. L., Threshold Average Precision ( $TAP-k$ ): A Measure of Retrieval Designed for Bioinformatics, *Bioinformatics Advance Access published on May 26, 2010*.

The Threshold Average Precision ( $TAP-k$ ) is  $MAP$  with a variable cutoff and terminal cutoff penalty.

For a single query the average precision ( $AP$ ) is computed by summing the precision at each rank that contains a true positive item and then dividing this sum by the number of positives ( $P$ ) for that query. If the retrieval system assigns to each retrieved item a score and the retrieved items are ranked in decreasing order of score, then it may be useful to cut off the retrieval at some fixed score level  $x$ . We can compute the average precision with cutoff  $x$  ( $APC_x$ ). This is the sum of the precision at each rank with a true positive item and a score  $\geq x$ , divided by the total number of positives for the query. Finally, suppose that  $y > x$  and further suppose there are no true positive items in the sum for  $APC_x$  that are below  $y$ . Then  $APC_y = APC_x$ . But clearly it would make more sense to choose the cutoff  $y$  than the cutoff  $x$ . To distinguish between these two cases we define the average precision with cutoff  $x$  and terminal penalty ( $APCP_x$ ). Let  $P_x$  be the precision at the last rank with score  $\geq x$  and let  $P$  be the total number of positives. Then define

$$APCP_x = \frac{P * APC_x + 1 * P_x}{P + 1}. \quad (1.1)$$

$APCP_x$  is just the weighted average of  $APC_x$  and  $P_x$  with most of the weight applied to  $APC_x$ , but  $P_x$  supplying the terminal penalty. In our hypothetical case  $P_y$  will be greater than  $P_x$  so that  $APCP_y$  is also greater than  $APCP_x$  and the score rewards the better choice of cutoff or equally penalizes the poorer choice. Whereas  $MAP$  is the average of  $AP$  over all the queries,  $TAP-k$  is the average of  $APCP_x$  over all the queries where  $x$  is chosen as the largest score that produces a median of  $k$  false positive retrievals over all the queries. The median is used here instead of the mean because it is more robust against noise and outliers.

There are some practical considerations when applying  $TAP-k$ . First, retrieval systems must produce scores commensurate with their rankings and these scores must be interpretable across different queries. Since most systems generate their retrieval by scoring this should not make the task any more difficult than usual. On the other hand some kind of score normalization may be necessary for some systems, depending on how the scores are constructed. An ideal score would be a probability estimate that the retrieved item is a true positive, but a score need not be a probability estimate for good performance. The score that is reported simply has to have the same implications for relevance of the item regardless of the query, for the best performance. Another important issue is the length of the retrieved lists returned by a system. If many of the retrieved lists are too short to have  $k$  false positives appear, then no cutoff score may produce a



11										
12										
13										
14										
15										

Here the cutoff score is 0.163 (the lowest score possible) and the  $APCP_5$  values are 0.583, 0.097, 0.125, 0, 0.333 and the average,  $TAP-5$ , of these numbers is 0.228. Here the  $TAP-5$  is lower than for example 1 because the system cut the retrieval off prematurely and this decreased the recall and thus the  $TAP-5$  score.

**Example 3.** Example 1 output again, but scores changed so they only reflect the rank and not the quality of the retrieved material.

	Q1 (5)		Q2 (5)		Q3 (5)		Q4 (3)		Q5 (5)	
	rel	score								
1	1	0.9	0	0.9	0	0.9	0	0.9	1	0.9
2	1	0.85	0	0.85	1	0.85	0	0.85	0	0.85
3	0	0.8	1	0.8	0	0.8	0	0.8	0	0.8
4	1	0.75	0	0.75	0	0.75	0	0.75	1	0.75
5	1	0.7	1	0.7	0	0.7	0	0.7	1	0.7
6	0	0.65	0	0.65	0	0.65	0	0.65	0	0.65
7	0	0.6	0	0.6	0	0.6	0	0.6	0	0.6
8	0	0.55	0	0.55	1	0.55	0	0.55	0	0.55
9	1	0.5	0	0.5	0	0.5	0	0.5	0	0.5
10	0	0.45	1	0.45	1	0.45	0	0.45	1	0.45
11	0	0.4	0	0.4	0	0.4	0	0.4	0	0.4
12	0	0.35	0	0.35	0	0.35	0	0.35	0	0.35
13	0	0.3	0	0.3	0	0.3	0	0.3	0	0.3
14	0	0.25	0	0.25	0	0.25	0	0.25	0	0.25
15	0	0.2	0	0.2	1	0.2	0	0.2	0	0.2

Here the scores no longer reflect quality and thus they do not give an accurate idea of where to cut off retrieval to obtain maximal efficiency. As a result there is a drop in  $TAP-5$  as compared with example 1. The cutoff score is 0.6 and the  $APCP_5$  values are 0.687, 0.170, 0.107, 0, 0.421 and the average,  $TAP-5$ , is 0.277.